Suggestibility Across Time: From False Memory to Deepfakes and Their Digital-Age Implications

Guodong Wu[1], Aron Elias[2], Jayln Huang[3#]

[1] University of Western Ontario
[2] Texas Academy of Mathematics and Science
[3] Irvine High School
[#] Advisor

ABSTRACT

This paper examines the malleability of human memory through the lens of false memory research, beginning with Loftus' seminal demonstrations of the misinformation effect. Drawing on both classic and contemporary findings, we explore how post-event information can distort memory at both the perceptual and conceptual levels, and how repeated exposure to misleading suggestions increases source confusion and produces vivid but inaccurate recollections. Real-world examples—including the McMartin Preschool Abuse Trial and Hillary Clinton's misremembered sniper-fire incident—highlight the pervasive impact of suggestibility beyond laboratory settings. Building on this foundation, we review how memory reconsolidation research reveals a critical period in which reactivated memories become temporarily destabilized and vulnerable to modification. Although this mechanism may pose ethical risks, it also presents opportunities for therapeutic interventions in conditions such as PTSD and substance use disorders. Replication studies of the original Loftus paradigm confirm the robustness of false memory implantation, even in larger and more ecologically valid samples. Recent work on AI-generated misinformation and deepfakes further underscores the urgency of this research, showing that large language models and synthetic media can dramatically increase false memory rates, particularly when people rely on intuitive (System 1) processing. Taken together, the evidence demonstrates that memory is highly reconstructive and susceptible to influence, emphasizing the importance of continued investigation and the development of strategies to enhance critical thinking and source monitoring in an increasingly digital world.

Keywords: *Misinformation, False Memory, Dual Theory, Deepfakes, Conceptual Elaboration, Memory Reconsolidation, Memory Malleability*

Introduction

Misinformation is a widely recognized term, commonly associated with false or inaccurate information. It can be introduced through media or interpersonal communication and, ultimately, may alter a person's memory or beliefs about events. In today's media-driven society, this phenomenon is known as the *misinformation effect*, a topic psychologists have studied extensively since the early 20th century. The misinformation effect refers to the presentation of post-event misinformation about a witnessed event, which can obscure, alter, or degrade the memory of the original experience (Schwartz, 2024, p. 393).

This phenomenon has been investigated by numerous psychologists, most notably Elizabeth Loftus, whose foundational work—such as the *"Lost in the Mall"* study and the stop/yield sign experiment—has demonstrated that false information introduced after an event can significantly distort memory (Loftus & Hoffman, 1989). These studies highlight the fallibility of human memory and underscore the unreliability of eyewitness testimony, which should not be solely relied upon in high-stakes legal contexts (Zaragoza et al., 2011).

Beyond post-event misinformation, additional research has revealed that memory can be altered through reconsolidation—a process whereby retrieved memories become malleable and susceptible to modification. When individuals recall a memory they believe to be accurate, the introduction of new or misleading information during that retrieval phase can change the original memory, potentially resulting in the formation of a new, distorted memory (Lee et al., 2017).

In the modern digital era, the rise of social media and the widespread circulation of unverified information have amplified the risks of misinformation. News outlets often prioritize being the first to publish a story, sometimes at the cost of accuracy or fact-checking. Moreover, because anyone with internet access can now publish content, false or misleading information can quickly reach a wide audience (Sherqulov, 2025). People often consume media passively, without critically evaluating its credibility. If a video appears professional or garners widespread attention, many viewers assume its authenticity. In some cases, even efforts to verify such content lead them to unreliable sources, further reinforcing false beliefs and contributing to the formation of false memories.

Although this paper does not focus on artificial intelligence (AI), it is worth noting the growing relevance of technologies such as deepfakes—realistic fabricated videos of public figures appearing to say or do things they never actually did. These AI-generated materials significantly increase the believability of misinformation, especially when perceived as coming from an authoritative source, rather than an unverified user or independent reporter (Chan et al., 2024).

This paper will examine a range of phenomena and scientific studies, including historical research dating back to the 20th century, that collectively underscore the significance of the misinformation effect. The power of this effect lies in its ability to demonstrate how exposure to false information can influence and alter memory. For instance, Loftus's *Lost in the Mall* study illustrates that entirely false autobiographical memories can be implanted, providing compelling evidence of the malleability of human memory.

Understanding how memories can be modified also has important clinical applications. In the context of treating Post-Traumatic Stress Disorder (PTSD), altering or desensitizing traumatic memories through techniques that utilize reconsolidation processes may provide therapeutic relief or even lead to the extinction of distressing stimuli (Lee et al., 2017).

In legal contexts, where eyewitness testimony can be a determining factor in verdicts, the implications of memory distortion are especially profound. If false information can alter someone's recollection—or even create memories of events that never occurred—then the justice system must proceed with caution. Preventing wrongful convictions hinges on understanding the factors that compromise memory accuracy.

In sum, this area of research holds immense societal value. Not only can false information distort existing memories, but entire fabricated experiences can be created. The goal of this paper is not to argue for a particular position or introduce new theories but rather to present current findings on the misinformation effect and explore its broader implications in a media-saturated, technologically evolving society.

Loftus' Original experiment

The original study by Loftus explores the concept of the *misinformation effect*, examining how an individual's memory can be altered or changed by the introduction of additional information. This suggests that a memory someone experienced can be modified, regardless of how strong the memory is or how recently it was formed. The rise of technology and social media outlets allows information to spread rapidly, meaning the impact of this research could help better prepare individuals and prevent misinformation from affecting them.

The importance of this research lies in its contribution to understanding memory errors and failures, which were previously considered forms of *interference*. Interference can occur in two forms: proactive interference and retroactive interference. Proactive interference (or proactive inhibition) occurs when memory for a current event is disrupted by earlier, similar memories. Retroactive interference (or retroactive inhibition) happens when memory for a past event is impaired by newer, similar events (Greene, 1992, p. 74). This means proactive interference occurs when older memories interfere with the current memory, such as when learning a new date is disrupted by remembering a previous one. Retroactive interference, on the other hand, which has been studied most extensively in psychology, refers to when new information interferes with the recall of earlier memories—such as learning a new date today causing someone to forget a previous one (Schwartz, 2024).

An example of retroactive interference was demonstrated in a laboratory experiment where participants watched a video of someone being killed in a crowd. Each group was shown the same video and received a written description of the event, but one group's description included tampered details. Although both groups saw the same video, the group exposed to the altered description integrated the false details into their memory of the event. This indicates that the added misinformation affected participants' recall of what they originally saw (Loftus & Pickrell, 1995).

Rather than explaining this solely as interference, researchers introduced the concept of the *misinformation effect*, where misleading information distorts memory. This led to the development of two main theories: the *trace impairment view* and the *coexistence hypothesis* (Schwartz, 2024). The trace impairment view proposes that the

original memory is altered by the misinformation, effectively replacing it with a new, inaccurate version. The coexistence hypothesis suggests that people retain both the original and the false memory and may retrieve either, depending on the context.

For example, if a person saw someone drink water and was then asked whether the person drank water or juice, followed by another question asking whether it was pop or beer, the individual might still say “water” or “juice” rather than choosing the false options. This implies they retained a copy of the original memory (Schwartz, 2024).

To further examine how false memories can be implanted, Loftus conducted the well-known *Lost in a Shopping Mall* experiment. The goal of this study was to implant a false memory into an individual’s mind—specifically, a childhood experience of getting lost in a shopping mall around the age of five. Researchers first gathered participants and enlisted help from their guardians or close relatives, such as mothers or siblings, to create four childhood memory narratives: three true events and one false event (being lost in the mall). The false event was described using details provided by the family to enhance believability.

There were 24 participants—three males and 21 females. Each participant was assigned to a student researcher who also worked with their family member. Participants received a booklet containing instructions and four stories. The third story in the booklet was always the fabricated "lost in the mall" event. Each event was about a paragraph long and followed by a blank space where participants could write down any additional details they recalled. If they did not remember anything, they were instructed to write, “I don’t remember.” An example of the fabricated memory used in the study is as follows:

“You, your mom, Tien, and Tuan all went to the Bremerton K-Mart. You must have been 5 years old at that time. Your mom gave each of you some money to get a blueberry Icee. You ran ahead to get into the line first and somehow lost your way in the store. Tien found you crying to an elderly Chinese woman. You three went together to get an Icee” (Loftus & Pickrell, 1995, p. 721).

In the first interview, researchers spoke with parents or siblings to collect verified events from the participant’s childhood (ages four to six). These events had to be factual—not folklore or secondhand stories—and could not be traumatic. The relative also had to provide details of a plausible shopping trip to support the implantation of the false memory.

The second interview occurred after participants completed the booklet and was conducted in person or over the phone. Participants were asked to recall the four events. If they struggled, they were given short sentences from the original narratives as memory cues. After each story, they were asked to rate the *clarity* of the memory on a scale from 1 to 10, and how likely they were to remember more details over time on a scale from 1 to 5.

The final interview followed a similar format, with participants again asked to recall the four events and rate them on the same two scales. At the end of this session, participants were debriefed: they were told the study’s purpose, asked to identify which event was false, and then informed of the correct answer along with the supporting evidence.

The researchers followed strict criteria to determine whether participants genuinely believed the false event. Participants who did not meet these criteria were excluded from the false memory analysis, ensuring the reliability of the results. The findings showed that 6 out of 24 participants (25%) accepted the false event as a real memory. Participants also used more words when describing true memories (M = 138.0) than false ones (M = 49.9) (see Figure 1).

For those who believed the false memory, researchers analyzed clarity and confidence ratings. In the first interview, the mean clarity score for true events was 6.3, compared to 2.8 for the false event. In the second interview, the mean confidence score was 6.3 for true events and 3.6 for the false event (see Figure 2). These results demonstrate that people can be led to believe in entirely fabricated experiences, especially when false details are introduced in a plausible and repetitive way.

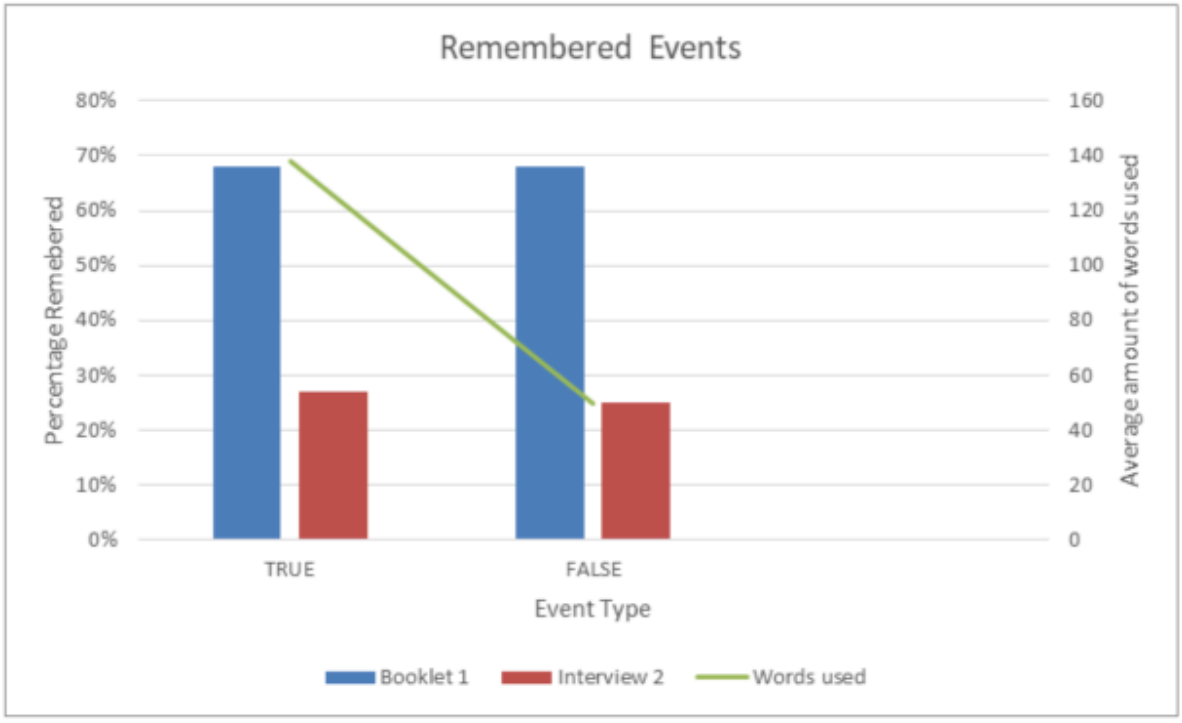

Figure 1. *Remembered Events. Twenty-four subjects were asked to remember the events from the booklet and second interview stages, and the amount of words used to recall each type of event are shown in the chart. Events show how many recall the memory and in which category they go.*

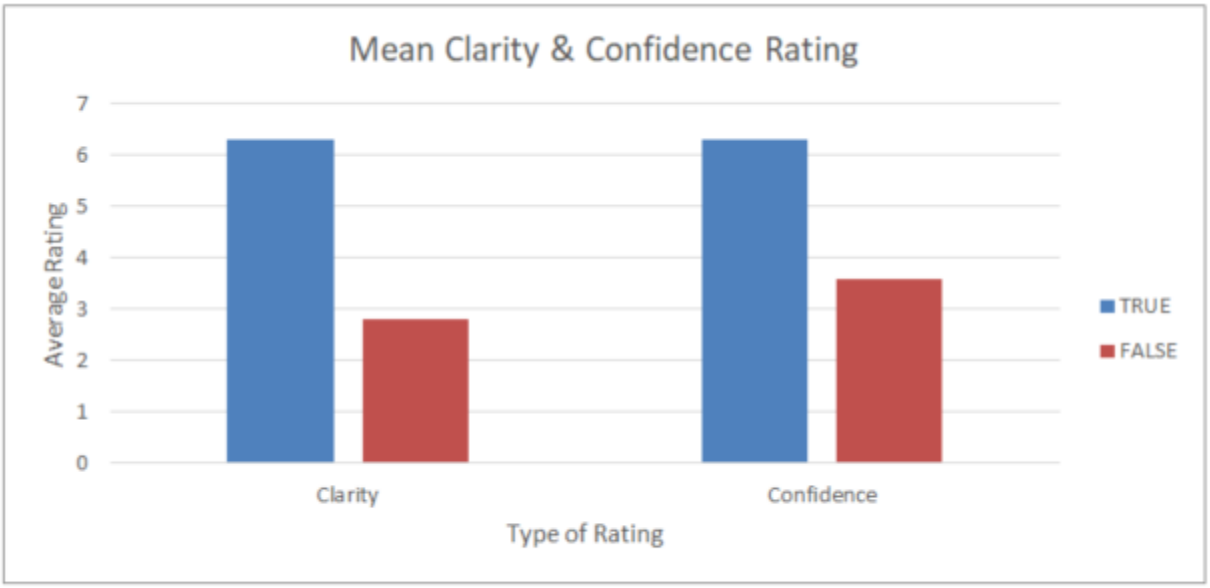

Figure 2. *Mean Clarity and Confidence Rating. Results showing ratings for both event types, calculated using the Interview 1 and Interview 2 ratings*

Some participants figured out which story was false through the process of elimination, but others continued to believe in the implanted memory. One participant, despite contacting her parents for confirmation that the event had not occurred, still had doubts about its falsehood.

Skeptics have argued that participants might have confused the false memory with actual past events, such as other times they were lost. However, the researchers observed how these false memories evolved from mere suggestions into seemingly authentic recollections. The misinformation, over time, became integrated with fragments of real experiences. This highlights how false memories can be constructed through suggestive techniques. While these may not always result in fully fabricated memories, misinformation can lead to distorted recollections that blend fiction with real or partial events.

Additional Research Showing False Memory Manipulation

Conceptual Elaboration

Research conducted by Zaragoza et al. has demonstrated the role of false memory in *conceptual elaboration* and *perceptual elaboration*. Perceptual elaboration refers to the processing of sensory information about an object's physical characteristics (e.g., shape, color), whereas conceptual elaboration involves understanding the object's meaning, function, or relationship to other concepts (Coutanche, 2021). Simply put, perceptual elaboration emphasizes how something looks, while conceptual elaboration focuses on what something is and how it functions.

Memory research has shown that life events are not static but are malleable and susceptible to contamination through both conceptual elaboration and external suggestion. Encouraging individuals to imagine fictitious events—particularly those rich in sensory and spatiotemporal detail—can increase false memory formation. This is primarily because mental imagery often mimics the qualities of authentic perceptual experiences (Zaragoza et al., 2011).

Conceptual elaboration has been shown to increase false memories for repeatedly suggested events. Repeated exposure to misleading suggestions significantly increases false memory rates, with the highest rates occurring when conceptual elaboration is involved. These findings suggest that while both perceptual and conceptual elaboration can contribute to false memory formation, conceptual elaboration—which involves deeper, meaning-based processing—has a particularly strong effect. Importantly, the rise in false memories is not merely the result of improved memory for the suggested content but is instead attributed to greater *source confusion*, lending support to theories such as the *Source Monitoring Framework* (SMF) (Zaragoza et al., 2011).

Conceptual elaboration has been found to enhance false remembering, especially when individuals are encouraged to link suggested details to meaningful, plausible episodes. When questions were grouped by episode (i.e., across scenes that form a coherent narrative), results replicated previous findings: conceptual elaboration produced the highest false memory rates (Zaragoza et al., 2011). This is likely because suggestions are more easily integrated into a cohesive narrative, which increases the plausibility of the false event and facilitates memory distortion.

The researchers argue that conceptual elaboration increases false memory by integrating suggested and witnessed events across meaningful dimensions—such as causal and thematic links—thus enhancing the *embeddedness* of the suggested information within the overall memory representation. This embeddedness makes suggested details phenomenologically similar to real memories, leading to misattribution errors.

Overall, the findings demonstrate that conceptual elaboration is a powerful and context-sensitive contributor to false memory formation. They reinforce the importance of how post-event information is organized and cognitively processed, both in controlled experimental settings and in real-world situations (Zaragoza et al., 2011).

False Memory and Misinformation Examples

Two prominent real-world examples of memory manipulation and false memories are the McMartin Preschool Abuse Trial and Hillary Clinton’s false memory of being "under sniper fire" during a visit to Bosnia. The McMartin trial, one of the longest and most expensive criminal cases in American history, illustrates how suggestive interviewing techniques can lead to false memory implantation, especially in children. It also highlights the risks of relying too heavily on single expert witnesses in legal proceedings (Douglas). In contrast, Hillary Clinton’s false memory was revealed during a speech at George Washington University in March 2008, where she recounted landing in Bosnia under sniper fire and having to duck for cover upon exiting the aircraft. However, video evidence and eyewitness testimony later contradicted this account, showing that the arrival was peaceful and without danger

(Kessler, 2016). Both cases underscore the malleability of memory and the importance of rigorous evidence and objective verification in both legal and political contexts.

The McMartin Criminal Case

The McMartin trial, which spanned over seven years and cost the government approximately $15 million, remains one of the longest and most expensive criminal cases in American history. It highlights several critical issues within the legal system, including flawed investigative procedures and the dangers of suggestive interviewing techniques, particularly when involving children.

The case began when a mother with a diagnosed mental disorder accused a staff member at the McMartin Preschool of sexually abusing her son. The situation quickly escalated, prompting an investigation that involved interviewing nearly 400 children. Many of these interviews employed improper and coercive methods, leading to widespread claims of ritual abuse. Despite the absence of physical evidence and emerging signs of false testimony, the trial proceeded (Douglas).

During the first trial, the prosecution relied heavily on child testimony and expert witnesses who supported the validity of the children's claims. However, defense attorneys raised significant concerns about the techniques used by therapist Kee MacFarlane, pointing to inconsistencies in testimony, suggestive questioning, and a lack of corroborating evidence. The case served as a clear example of how misinformation and poor interviewing methods can implant false memories, particularly in vulnerable individuals such as children. Many of the children eventually came to believe in the stories constructed during these interviews. Although some of the accused were acquitted, others faced retrials, extending the trial's duration and further exhausting public resources (Douglas).

The case exposed systemic failures in how allegations of child abuse were handled, especially the use of suggestive forensic interview techniques and the influence of public hysteria on legal proceedings. These concerns underscore the rationale behind judicial instructions that jurors avoid media coverage related to ongoing trials (Douglas).

Years later, some of the children involved in the case came forward in interviews and admitted to fabricating their stories. They explained that they felt pressured to conform to the group and that the interviewers' suggestive questioning shaped their responses. This further reinforced the dangers of leading questions when attempting to recover or reconstruct repressed memories. For example, one interviewer reportedly asked, "Can you remember the naked pictures?"—a question that presupposes the existence of such photos and subtly implies the child should recall them. This kind of framing leads children to form *false memories*, especially when reinforced over time through repeated questioning and subtle suggestions (Kettler, 2024).

The consequences of this flawed process were devastating. An innocent man spent years in prison for a crime he did not commit, and millions of dollars were spent on a trial that many now agree was unwarranted. The McMartin case has since become a cautionary tale about the perils of moral panic, prosecutorial overreach, uncritical media amplification, and flawed investigative practices. It illustrates the urgent need for evidence-based forensic procedures, judicial restraint, and responsible media coverage in cases involving child testimony and allegations of abuse (Douglas).

Hillary Clinton's Bosnia's Trip

In 2008, Hillary Clinton claimed that she had come under sniper fire upon her arrival in Bosnia in March 1996. Clinton's recollection of events diverged significantly from the documented reality. Her statement—that she and her delegation had to run with their heads down under threat—was contradicted by contemporaneous news reports, video footage, photographs, and firsthand accounts (Kessler, 2016). Journalistic and official records depict a calm and orderly arrival ceremony at Tuzla Air Base, during which Clinton was greeted by local dignitaries and a young girl who presented her with flowers. Reporters on the scene, including John Pomfret, as well as military personnel such as Maj. Gen. William Nash confirmed that there was no active threat or gunfire during the visit.

Clinton appeared to genuinely believe her account, suggesting that the false memory felt real to her. Her recollection may have been influenced by prior briefings about potential (but unrealized) sniper threats. Combined with the stress of being in a conflict zone and repeated exposure to what might have happened, her memory integrated these suggestions into the actual event, despite no sniper activity taking place. This incident underscores the strength of the misinformation effect and the power of suggestibility in shaping memory.

Today, this episode is frequently cited as a case study in political embellishment, the fallibility of memory, and the lasting effects of public misstatements. It also underscores the critical importance of journalistic

fact-checking and highlights how inaccurate recollections—once made public—can affect political credibility, erode public trust, and demonstrate the malleability of memory (Kessler, 2016).

Memory Reconsolidation

As previously discussed, memory is a dynamic and malleable process. To reduce errors and strengthen core memories, researchers have focused on the processes of memory consolidation and reconsolidation. Memory consolidation is a neurological process through which memory traces become stabilized and permanently stored in long-term memory (Schwartz, 2024). Memory reconsolidation, by contrast, refers to the process whereby a previously consolidated memory, once reactivated, becomes temporarily destabilized and susceptible to modification before it is restabilized (Lee et al., 2017). In simpler terms, consolidation strengthens a memory for long-term storage, whereas reconsolidation allows a reactivated memory to be updated, similar to editing and re-saving a document.

False memories can emerge during the reconsolidation window, a period in which long-term memories become unstable and open to change before being re-stored. However, the reconsolidation model is not universally accepted. Some researchers have proposed alternative explanations for the mismatch between memory encoding and retrieval, such as state-dependent learning and uninhibited extinction. In state-dependent learning, amnesia is attributed to a mismatch between the physiological or psychological states during encoding and retrieval. In the uninhibited extinction framework, reactivation-dependent amnesia is interpreted as a form of extinction learning that lacks inhibitory control, making the suppressed memory more vulnerable to later reactivation (Lee et al., 2017).

Despite ongoing debate, research into memory reconsolidation has led to promising applications, particularly in the treatment of psychiatric disorders such as PTSD, phobias, and substance-related addictions, including heroin and nicotine dependency. Understanding how memories can be modified has sparked concern about the potential for abuse, such as the implantation of false memories by therapists or others, which could later be falsely recalled as evidence in legal contexts. On the other hand, this understanding offers hope for healing, providing individuals with the tools to reframe traumatic memories or weaken maladaptive emotional responses.

Overall, while concerns and debates persist, reconsolidation research has identified two critical factors necessary for successful memory modification: memory destabilization and effective disruption. Memory destabilization occurs following reactivation and typically involves a prediction error, or a mismatch between expected and actual outcomes. Effective disruption or modification must then occur during the reconsolidation window, whether through pharmacological intervention or behavioral manipulation (Lee et al., 2017). These findings highlight both the promise and complexity of altering human memory, reinforcing the need for ethical safeguards and continued research.

Heroin/Nicotine Addiction

Many studies on treating drug addiction have focused on pharmacological interventions. However, these approaches face limitations due to human safety concerns. Traditional cue-exposure therapies have also shown limited success, largely because of phenomena such as reinstatement, renewal, and spontaneous recovery, all of which can re-trigger drug-seeking behavior. A more recent line of research has explored the application of reconsolidation theory, which involves reactivating drug-associated memories followed by extinction training, to weaken or disrupt the original memory traces and reduce sensitivity to cues (Xue et al., 2012).

In a series of preclinical and clinical experiments, researchers examined whether exposing rats and humans to drug-associated cues—effectively reactivating the memory of the drug's effects—before extinction training (either 10 minutes, 1 hour, or 6 hours later) would interfere with memory reconsolidation and thereby reduce drug-seeking behaviors and cravings. Results demonstrated that extinction training conducted within the reconsolidation window (i.e., 10 minutes or 1 hour after memory retrieval) led to significant reductions in reinstatement, spontaneous recovery, and renewal of drug-seeking behavior in rats. In contrast, a 6-hour delay between memory retrieval and extinction did not yield these effects, suggesting that the procedure is highly time-sensitive (Xue et al., 2012).

In rat models, the memory retrieval-extinction procedure successfully reduced cocaine- and morphine-conditioned place preference (CPP), as well as drug-primed reinstatement and spontaneous recovery in both CPP and self-administration paradigms. The conditioned place preference model is a widely used behavioral assay to investigate the association between context and reward, including both natural rewards and drugs of abuse (McKendrick & Graziane, 2020). Notably, this behavioral outcome was accompanied by molecular changes in the

brain: there was increased expression of PKMζ (Protein Kinase M Zeta) in the infralimbic cortex and decreased expression in the basolateral amygdala—brain regions known to be involved in memory maintenance and extinction (Patel & Zamani, 2021).

In clinical trials involving human participants with heroin addiction, similar effects were observed. When extinction training was followed by memory retrieval just 10 minutes later, participants exhibited reduced cue-induced heroin craving and lower blood pressure responses. These effects were not present when extinction occurred after a 6-hour delay, reinforcing the importance of the reconsolidation window. The human findings parallel the time-dependent effects seen in animal models.

Although the exact mechanisms underlying these effects are still being investigated, the current evidence supports a dual-process model involving both disruption of reconsolidation and enhancement of extinction memory. These processes appear to be mediated, in part, by PKMζ-dependent plasticity in key brain regions associated with emotion and memory regulation. Importantly, this memory retrieval-extinction procedure offers a promising, non-invasive, and behavior-based adjunct to traditional addiction treatment. Unlike pharmacological interventions, it does not rely on medication and may reduce the risk of relapse by directly weakening the memory associations that drive compulsive drug use.

Anxiety and Trauma-related Disorders

Traditional extinction-based posttraumatic stress disorder (PTSD) treatments (e.g., exposure therapy) demonstrate variable and often temporary effectiveness. These treatments are vulnerable to relapse phenomena such as spontaneous recovery, contextual renewal, reinstatement, and rapid reacquisition, suggesting that extinction suppresses rather than erases traumatic memories. Utilizing memory reconsolidation mechanisms, the visual-kinesthetic dissociation (V/KD) protocol is proposed as a non-pharmacological treatment designed to transform traumatic memories (Gray & Liotta, 2012). The technique involves brief memory activation followed by dissociative visualization—such as watching the trauma like a movie in black and white and in reverse—to prevent retraumatization and facilitate memory updating.

During the procedure, the individual imagines themselves watching a "movie" of their trauma from multiple levels of detachment. The psychologist then works backward from the re-experiencing of the traumatic memory and begins and ends the session in a safe emotional state to ensure psychological stability before and after the dissociative process. In contrast to extinction-based treatments, reconsolidation aims to update or even erase original memory traces, thereby reducing symptom intensity and avoiding relapse phenomena such as spontaneous recovery, reinstatement, or contextual renewal.

Gray and Liotta (2012) examined a case study involving an individual who participated in three one-hour videotaped sessions, with a three-day interval between sessions. The PTSD Checklist—Civilian Version (PCL-C) was administered before treatment, at the beginning of the second session, after the third session, and again 30 days post-treatment. The participant's pre-treatment score was approximately 90%. Following the first session, the score decreased to 30%. After the third session, all symptoms had disappeared. At the 30-day follow-up, symptom scores remained at zero, and the client reported no recurring symptoms.

The V/KD protocol presents a theoretically grounded, effective, short-term, and non-traumatizing intervention for PTSD by leveraging reconsolidation rather than extinction mechanisms. While the initial findings are promising, the approach currently relies on anecdotal evidence, underscoring the need for rigorous empirical validation in future research (Gray & Liotta, 2012).

Large Loftus Replication Study

Scientific knowledge is built through experimentation and replication, which brings into question the replicability of Loftus and Pickrell's false memory study. Numerous similar studies have demonstrated that memory can be manipulated or altered by introducing new information. These findings show that it is indeed possible to implant false memories in individuals.

One such replication study, conducted in Ireland with a larger sample size, followed the same procedures as the original Loftus and Pickrell study and had an additional group to explore legal implications. This group was included to examine how others might perceive a false memory as genuine, especially in legal contexts. The replication study successfully reproduced the original results: participants developed false memories of being lost in

a mall. Interestingly, the rate of false memory formation was even higher in this study, possibly due to the increased sample size, suggesting that a larger sample could further elevate this percentage.

The significance of this replication lies in its methodological rigor. The researchers reproduced the procedures, interview structure, booklet format, and even included the option for participants to withdraw mid-study. In addition, they incorporated phenomenological questions during the second interview, requiring participants to recount the memory and provide further details.

To ensure accuracy and objectivity, two observers recorded participants' responses during interviews. They used Cohen's Kappa, a statistical measure of inter-rater reliability, to assess agreement between observers. In cases of discrepancy, a third evaluator—typically a more experienced supervisor—reviewed the data to resolve inconsistencies.

An additional and innovative component of the replication study involved presenting recordings of participants recalling their false memories to a mock jury. Twelve jurors listened to audio recordings and read transcripts of the second interview. They were informed only that the study aimed to investigate how people perceive memories shared by others. After reviewing the material, jurors answered three yes-or-no questions:

- Do you believe this event really happened?
- Is this person describing something they remember happening?
- Is this person describing something they believe happened?

At the end of the session, jurors were asked to rate the clarity, detail, emotional intensity, and plausibility of the memory on a scale from 1 (not at all) to 100 (extremely).

This element of the study provided important insights into how memory is perceived in legal contexts. In judicial systems, a group of jurors must evaluate the credibility of someone's memory to make decisions that can result in life-altering outcomes. If a false memory is presented in a way that appears believable, it could lead to wrongful convictions or allow guilty individuals to go free.

Overall, the study successfully replicated Loftus and Pickrell's (1995) findings, showing that false memories—such as being lost in a shopping mall—can be implanted through suggestion. While some critics argue that the memory of being lost in a mall is too commonplace to be considered truly "false," the study demonstrates that even mundane memories can be convincingly fabricated. Moreover, other research has shown that even implausible events can be implanted, suggesting that while event plausibility plays a role, it is not the only factor that determines memory implantation success.

One limitation of this study is that getting lost in a mall is a relatively common experience, and some participants may have recalled genuine events mistakenly denied by their informants. Nonetheless, the findings reinforce the conclusion that false memories can be vivid, detailed, and emotionally resonant—both for the individuals experiencing them and for those hearing the accounts. This has important implications for psychology, the legal system, and our broader understanding of the malleability of human memory.

Apposing Research to False Memory

Research conducted by Andrews and Brewin focused on critiquing the replication study by Murphy et al. and raised questions about the findings related to false memory implantation and the misinformation effect. The researchers examined how false memory implantation truly works and highlighted two critical questions:

1. Whether participants who were deemed to have developed false memories reported the six "core details of the Lost in the Mall story."
2. Whether participants' false memory reports contained real episodic details.

The second question directly challenges the assumptions of the implantation paradigm, which presumes that such events never occurred and, therefore, any subsequently produced memory must be false. Andrews and Brewin argue that no actual misinformation may have occurred, but rather that a misclassification of real memories as false memories took place. Simply put, fragments of genuine experiences—such as visiting a common store at a certain age—may be mistakenly integrated into a fabricated narrative of being lost.

Applying their exclusion criteria, Andrews and Brewin excluded participants who failed to explicitly recall all six core details of the implanted event, those who self-reported not remembering the event, and those whose accounts relied on potentially real memories or details provided to them. After these exclusions, they concluded that only about 4% of participants had genuinely developed false memories.

However, Andrews and Brewin's proposition—that false memories must consist entirely of fabricated details—contradicts a substantial body of empirical evidence. Rather than requiring total falsification, researchers should consider the extent to which a memory includes false elements. In legal settings, for instance, if a witness

misremembers a small detail, their entire testimony should not be dismissed outright; rather, it should be evaluated critically and contextually.

Regardless of how the findings are interpreted, other research has shown that even veridical autobiographical memories—that is, confidently recalled memories of real past experiences—can contain inconsistencies or errors. Studies have demonstrated that reports of genuine emotional experiences can change significantly over time, with specific details becoming distorted, exaggerated, or disappearing altogether. Nevertheless, individuals continue to recall the core or "essence" of the experience (Wade et al., 2025).

Finally, even after applying strict exclusion criteria, a percentage of participants still demonstrated signs of false memory implantation. This outcome reinforces the possibility that false memories can occur. With a larger and more diverse sample size, this percentage may even increase. In scientific research, the ability to produce measurable results—even under scrutiny—should not be disregarded but instead considered a meaningful contribution to the ongoing study of human memory.

AI Research on False Memory

Large Language Model (LLM)

As artificial intelligence (AI) becomes increasingly integrated into daily life—serving roles ranging from virtual assistants to memory aids—it raises important concerns about its cognitive effects, particularly its potential to introduce or reinforce misinformation. Previous research has shown that humans are susceptible to false memories introduced through misleading questions or social cues. However, the specific impact of large language models (LLM)-based conversational AI on false memory formation has remained largely unexplored (Sherqulov, 2025).

LLMs are a class of foundation models trained on immense amounts of data, enabling them to understand and generate natural language and other types of content across a wide range of tasks. These models provide foundational capabilities that power numerous applications and use cases, including answering questions, summarizing information, and generating creative content (Caballar, 2023).

A recent two-phase experimental study explored this phenomenon by involving 200 participants who watched a CCTV video of a robbery. Participants were randomly assigned to one of four experimental conditions: (1) a control group, (2) a survey-based misinformation group, (3) a pre-scripted chatbot group, and (4) a generative chatbot group using a conversational LLM that provided confirmatory and elaborative feedback. The study simulated real-world forensic scenarios—such as AI-assisted witness interviews—and assessed both immediate and delayed (one week later) effects on false memory formation (Chan et al., 2024).

The generative chatbot condition led to the highest rate of false memories, with 36.4% of participants reporting details that never occurred. This was nearly three times the rate observed in the control group (15.2%) and significantly higher than in the survey-based (21.6%) and pre-scripted chatbot (26.8%) groups. These differences were statistically significant. Moreover, all intervention groups reported increased confidence in their false memories compared to the control group, with the generative chatbot condition yielding the highest levels of confidence (Chan et al., 2024; see Figure 3).

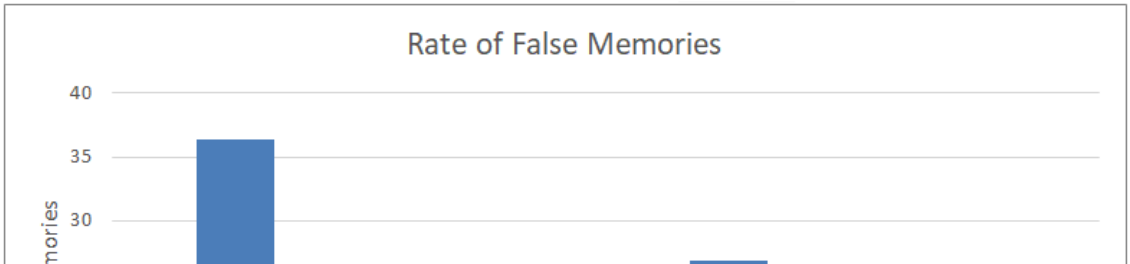

Figure 3. *Rate of False Memories. Results of the different conditional groups and their likelihood of farming false memories, all found to be statically significant*

Researchers identified four major factors that increase susceptibility to AI-induced false memories: age, cognitive traits, environmental influences, and emotional state. In terms of age, older adults—particularly those over the age of 60—had a 47.2% accuracy rate in identifying whether an event was AI-generated. In contrast, adults aged 18–39 had a significantly higher accuracy rate of 79.6%, nearly double that of the older group (Sherqulov, 2025).

Regarding cognitive traits, individuals with lower critical thinking skills and limited familiarity with chatbots were more prone to accepting false information as true. This vulnerability likely stems from a reduced ability to discern whether the information source was human or machine-generated. Environmental influences also played a role, as individuals with adverse early-life conditions and weak social bonds showed elevated risk of AI-induced false memories. Finally, emotional state affected susceptibility; individuals with chronic negative moods—such as those associated with depression—demonstrated greater vulnerability to false recall (Sherqulov, 2025).

.

Deep Fakes and Memory Malleability

Fake news exploits the dynamics of social media and weakens the authority of traditional journalism. It exhibits strong virality and psychological influence, even among educated and rational individuals. As a result, considerable research has been devoted to understanding why fake news is so compelling and widely disseminated, even when its veracity is questionable. These findings suggest that the root of the problem may lie in cognitive biases and identity-driven belief systems that override critical thinking.

One prominent explanation for this phenomenon is Identity Protective Cognition Theory (IPCT). According to this theory, individuals—regardless of their educational level—are more likely to accept information that aligns with and affirms their social or political identities (Liv & Greenbaum, 2020). Rather than evaluating information solely based on factual accuracy, people are motivated to preserve their group affiliations and worldview, which makes them more susceptible to misinformation that reinforces these beliefs.

Another relevant framework is the dual-process theory, which posits that human cognition operates through two systems: System 1 and System 2. System 1 is fast, automatic, and emotionally driven, while System 2 is slower, deliberative, and logical (Minda, 2021). Although both systems are active, System 1 dominates most daily decision-making due to its efficiency and low cognitive cost. Evolutionarily, this system developed to support rapid responses to immediate threats, relying on heuristics and intuitive judgments. In contrast, System 2 is capable of complex reasoning but requires more mental energy, making it less frequently activated, especially during passive activities like scrolling through social media (Minda, 2021).

Fake news exploits the cognitive shortcuts embedded in System 1, misleading individuals even when they are capable of rational analysis. While System 2 can intervene, doing so is cognitively taxing and inefficient for routine tasks, including idle media consumption. As a result, even informed individuals can fall prey to misinformation when cognitive resources are low or when emotional content bypasses analytical scrutiny.

A striking example of how false memories can influence large populations is the Mandela Effect—a phenomenon in which large groups of people collectively misremember events, phrases, or details. These memory distortions often arise from social reinforcement, linguistic distortion, and repetition of misinformation, all of which highlight the malleable and reconstructive nature of human memory.

Deep Fakes and False Memories

Deepfakes are becoming increasingly sophisticated due to rapid technological advancements, raising serious concerns about their potential to influence memory and perception. Researchers have investigated whether high-quality deepfake videos are more effective than text or image formats in generating false memories of fabricated news stories. Participants in one study rated high-quality deepfake videos as more realistic and more likely to convince others (Murphy & Flynn, 2021). However, statistical analyses revealed a more nuanced picture, indicating that deepfake videos—regardless of their quality—did not significantly increase false memory formation compared to other formats.

Murphy and Flynn (2021) also found that the presentation format—whether video, photo, or text—did not significantly influence participants' belief in or memory of the stories once they were cautioned about the potential for misinformation. These results support the notion that the existence of deepfake technology does not inherently

make misinformation more memorable. This finding corroborates prior research by Nash (2018), which showed that even poorly manipulated media can distort memory.

The study's findings suggest that deepfake videos do not consistently enhance false memory creation and that their impact may be context-dependent. These results align with earlier studies highlighting the power of narrative-based misinformation and with research indicating that while photographs and videos can increase a story's plausibility or familiarity, they may reduce source-monitoring errors. This is because highly detailed visual formats can limit the need for imaginative elaboration—a process closely linked to false memory development (Garry & Wade, 2005; Johnson et al., 1993).

Nonetheless, the malleability of memory and the compelling realism of deepfakes suggest that their psychological effects warrant further exploration. While current evidence does not indicate a uniform risk, the potential for deepfakes to alter perception and memory—especially in emotionally charged or politically sensitive contexts—remains an important avenue for future research.

Limitations of Research

Research in the field of memory—particularly regarding false memories—has been ongoing for decades and plays a critical role in both the treatment of memory-related disorders and the protection against memory manipulation. One common limitation across this body of literature is the difficulty in reliably identifying false memories and accurately classifying them. Participants often conclude that a false memory is true, while others clearly remember real events that conflict with the implanted ones. Many studies that challenge the validity of false memory findings often point to alternative memory phenomena—such as source misattribution and memory errors—as explanations for these results.

Another significant limitation involves the verification of event authenticity. To determine whether a memory is truly false, researchers must often rely on the accounts of relatives or close acquaintances regarding the participant's life history. However, memory is inherently malleable and subjective. It is entirely possible that informants misreport an event—either claiming it occurred when it did not or vice versa—leading to misclassification of genuine memories as false or false memories as real. While event plausibility plays a role in memory implantation, prior research has shown that even less plausible or uncommon events can be implanted successfully, suggesting that plausibility alone does not determine the success of false memory formation.

Within the domain of AI-related memory research, a methodological limitation is the frequent use of single-exposure designs, which do not accurately reflect real-world conditions. In everyday life, individuals are often repeatedly exposed to the same media content. This repeated exposure can increase perceived truthfulness due to the illusory truth effect, where familiarity is mistaken for accuracy. Moreover, previous findings suggest that repetition can even increase the ethical acceptability of spreading misinformation, regardless of whether it is believed. These observations point to important directions for future research, particularly the need to examine how repeated AI-driven misinformation may affect memory, belief, and behavior in more ecologically valid contexts.

Future Approaches for this Research

A prominent future direction in false memory research involves the integration of neuroscientific methods to identify the neural mechanisms that underlie memory distortion. Advanced neuroimaging tools—such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG)—hold promise for distinguishing the neural signatures of true versus false memories, particularly in brain regions associated with memory consolidation, source monitoring, and emotional regulation. In parallel, non-invasive brain stimulation techniques such as transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) may be employed to causally probe, and potentially modulate, susceptibility to memory distortion.

Future research should also prioritize understanding individual differences in vulnerability to false memories. This includes exploring cognitive variables such as working memory capacity and attentional control, as well as personality traits, emotional states, and neurodevelopmental or neurodegenerative conditions. Longitudinal research designs will be particularly valuable in tracking the formation, stability, and evolution of false memories across the human lifespan, from childhood through older adulthood.

Additionally, developing and testing intervention strategies aimed at reducing susceptibility to false memories will be critical. This includes evaluating the effectiveness of source-monitoring training, media literacy

programs, and encoding-strengthening techniques. Furthermore, the influence of factors such as sleep, stress, and affective states on memory reliability warrants further empirical investigation.

Conclusion

The foundational work of Elizabeth Loftus marked a major turning point in memory research by demonstrating that recollections are not static records of past experience but can be reshaped through exposure to misleading information. Her original studies introduced the concept of the misinformation effect and showed how memory errors and failures are often the result of post-event suggestions. Since then, research has expanded into related areas such as conceptual elaboration, which explains how false memories become more believable when individuals integrate misleading details into meaningful narratives.

Real-world examples such as the McMartin Preschool Trial and Hillary Clinton's misremembered experience of coming under sniper fire demonstrate that false memories are not limited to laboratory settings. In both cases, suggestive questioning and plausible contextual information resulted in the creation of vivid yet inaccurate recollections. These examples show that even well-intentioned memory reconstructions—such as preparing for potential threats—can blend with real experiences and lead to confident, but false, memories.

Importantly, the process of reconsolidation suggests that memories become temporarily malleable each time they are retrieved, making them vulnerable to modification. This malleability has both risks and benefits: while misinformation can corrupt memory during this window, carefully timed interventions can also be used to weaken maladaptive memories. Indeed, reconsolidation-based techniques are showing promise in the treatment of substance use disorders and trauma-related conditions by reducing the emotional strength of harmful memories.

This research is particularly relevant in contemporary contexts characterized by rapid information dissemination, social media use, and the increasing presence of AI-driven technologies such as large language models. Recent studies show that AI-based conversational agents and deepfakes can exacerbate false memory formation, especially among vulnerable populations and individuals with lower critical thinking skills. Repeated exposure to AI-generated misinformation may increase perceived truthfulness and confidence—even when the information is incorrect—raising serious ethical and social concerns. As people increasingly rely on AI-generated information in everyday settings, they may unknowingly integrate inaccuracies into their memories—especially when processing information quickly using intuitive, System 1 thinking. This underscores the importance of continued research, replication of foundational studies, and the development of educational strategies aimed at improving critical thinking and source monitoring in an increasingly digital world.

References

Caballar, R. D. (2023, November 2). *What are large language models (llms)?*. IBM. https://www.ibm.com/think/topics/large-language-models

Chan, S., Pataranutaporn, P., Suri, A., Zulfikar, W., Maes, P., & Loftus, E. F. (2024). Conversational AI powered by large language models amplifies false memories in witness interviews. *CrimRxiv*. https://doi.org/10.21428/cb6ab371.6ae390a8

Coutanche, M. (2021). *The link between conceptual and Perceptual Information ...* OSFHOME. https://osf.io/e75ks/download

Douglas, L. (n.d.). *The McMartin Preschool Abuse Trial: An Account* . UMKC School of Law. https://famous-trials.com/mcmartin/902-home

Gray, R. M., & Liotta, R. F. (2012). PTSD: Extinction, reconsolidation, and the visual-kinesthetic dissociation protocol. *Traumatology*, *18*(2), 3–16. https://doi.org/10.1177/1534765611431835

Greene, R. L. (1992). *Human memory: Paradigms and paradoxes*. Psychology Press. 2025, https://doi.org/10.4324/9781315807195

Kessler, G. (2016, May 23). *Recalling Hillary Clinton's claim of 'landing under sniper fire' in bosnia - The Washington Post*. The Washington Post. https://www.washingtonpost.com/news/fact-checker/wp/2016/05/23/recalling-hillary-clintons-claim-of-landing-under-sniper-fire-in-bosnia/

Kettler, S. (2024, April 29). *The mcmartin preschool case: Satanic panic and child sexual abuse allegations*. A&E. https://www.aetv.com/real-crime/mcmartin-preschool

Liv, N., & Greenbaum, D. (2020). Deep fakes and memory malleability: False memories in the service of fake news. *AJOB Neuroscience*, *11*(2), 96–104. https://doi.org/10.1080/21507740.2020.1740351

Loftus, E. F., & Hoffman, H. G. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General*, *118*(1), 100–104. https://doi.org/10.1037//0096-3445.118.1.100

Loftus, E. F., & Pickrell, J. E. (1995). The formation of False Memories. *Psychiatric Annals*, *25*(12), 720–725. https://doi.org/10.3928/0048-5713-19951201-07

MacAskill, E. (2008, March 25). *Clinton forced to admit she exaggerated tale of Bosnian sniper fire*. The Guardian. https://www.theguardian.com/world/2008/mar/25/hillaryclinton.uselections2008

McKendrick, G., & Graziane, N. M. (2020). Drug-induced conditioned place preference and its practical use in substance use disorder research. *Frontiers in Behavioral Neuroscience*, *14*. https://doi.org/10.3389/fnbeh.2020.582147

Minda, J. P. (2021). *The Psychology of Thinking: Reasoning, decision-making and problem-solving*. SAGE Publications.

Murphy, G., & Flynn, E. (2021). Deepfake false memories. *Memory*, *30*(4), 480–492. https://doi.org/10.1080/09658211.2021.1919715

Patel, H., & Zamani, R. (2021). The role of PKMZ in the maintenance of long-term memory: A Review. *Reviews in the Neurosciences*, *32*(5), 481–494. https://doi.org/10.1515/revneuro-2020-0105

Schwartz, B. L. (2024). *Memory: Foundations and applications*. Sage. 2025, https://collegepublishing.sagepub.com/products/memory-5-287857

Sherqulov, I. (2025). *Ai-Induced False Memories: New Research Shows 87% Success Rate in Memory Manipulation*. https://doi.org/10.2139/ssrn.5142397

Wade, K. A., Riesthuis, P., Bücken, C., Otgaar, H., & Loftus, E. F. (2025). Still lost in the mall—false memories happen and that's what matters. *Applied Cognitive Psychology*, *39*(1). https://doi.org/10.1002/acp.70028

Xue, Y.-X., Luo, Y.-X., Wu, P., Shi, H.-S., Xue, L.-F., Chen, C., Zhu, W.-L., Ding, Z.-B., Bao, Y., Shi, J., Epstein, D. H., Shaham, Y., & Lu, L. (2012). A memory retrieval-extinction procedure to prevent drug craving and relapse. *Science*, *336*(6078), 241–245. https://doi.org/10.1126/science.1215070

Zaragoza, M. S., Mitchell, K. J., Payment, K., & Drivdahl, S. (2011). False memories for suggestions: The impact of conceptual elaboration. *Journal of Memory and Language*, *64*(1), 18–31. https://doi.org/10.1016/j.jml.2010.09.004